微博内向型传导热点发现与预测算法研究*

■ 王林森[1] 王学义[2]

[1] 西南医科大学人文与管理学院 泸州 646000 [2] 西南财经大学中国西部经济研究中心 成都 611130

摘要:[目的/意义]为应对微博内向型传导热点生灭速度快、热点特征不明显等问题,研究新型的微博内向型传导热点发现与预测算法。[方法/过程]针对上述问题,基于复杂网络分析方法,构建微博传导热点预测算法,该算法通过复杂网络节点模型扩展生成微博传导节点模型,发现内向型传导节点的传导子网;通过对传导节点序列实施热功率计算,对其信息传导覆盖范围以及未来影响力进行预测,进行传导热点发现及预测。[结果/结论]数据实验表明,该算法较之目前常用的热点预测算法,具有较高的传导热点覆盖率和准确率,且耗时较少,性价比较高。

关键词: 舆情监测 微博 复杂网络 传导节点 热点预测

分类号: G203 TP311

DOI:10.13266/j.issn.0252-3116.2018.03.009

微博作为现代网络媒体最具代表性的信息传播平台,以其信息传播迅速、信息内容丰富等特点受到了广大网民的青睐,因此成为舆情管理的重点监测对象之一[1]。与既往的网络媒体不同,微博系统一方面具有相对独立性,各个微博系统自生(本地原创)内容相当丰富,例如:国内四大微博系统(新浪、腾讯、搜狐、网易)日均的自生信息数量早已突破千万级大关;另一方面各微博系统间的信息传导活动非常频繁,系统互动密集,复杂网络理论中的传导(转帖或转载)作用显著,例如:《2017 年第一季度微博传播研究报告》表明:四大微博系统中超过 65% 的信息为非原创,其中超过 70% 的信息转载自其他传播系统,而其中 90% 以上的热点微博信息来自于占总数不足 2% 的传导节点。目前的舆情监测工作以及微博热点挖掘研究,主要专注于单一微博系统中的内容及信息发布者的热度识别与检测,而将传导节点,特别是传导热点,与自生型信息热点混为一谈,忽略了复杂网络研究的重要定论,即"外部信息传导与内部信息生成具有同等重要的意义,但需要区别对待",进而导致了热点监测不全面、不准确,舆情控制启动速度慢,关键控制点缺失等问题[2]。基于上述现象及当前研究存在的问题,本研究以复杂网络理论为工具,通过对微博网络节点进行关系建模,提出了一种微博传导热点的发现与预测算法(Information Pass & transmission Inward Node detecting and irognosis, IPIN)。需要指出的是:本研究主要针对内向型传导节点展开,即信息由外部网络世界,通过传导节点向微博系统内部传导扩散,下文中如不作特殊说明,"传导节点/热点"等同于"内向型传导节点/热点"。

1 前人工作与解决思路

首先是复杂网络传导节点理论运用于微博系统的模型构建问题。S. Aparicio 等人证明了微博系统具有复杂网络的典型特征,并指出传导节点与复杂网络中的信息路由尽管较为相似,但存在信息传播方式、主动被动运行等方面的差别[3]。M. Coletto 等人指出复杂网络中的传导节点较之微博系统中的传导节点更为复杂,可以抽取其中部分属性与操作,用以构建微博传导节点模型[4]。W. Maharani 和 C. Chelmis 等人通过复杂网络传导节点模型,扩展衍生出一个初步的内向型微博传导节点模型,并通过仿真实验证明此类节点对于微博信息互动的重要性[5-6]。但上述工作普遍存在仅有设计思路或总体方案,缺乏详细建模、处理算法与模型应用等问题。针对上述问题,本研究将以复杂网络传导节点的基本属性对微博传导节点进行描述和建

* 本文系国家社会科学基金青年项目"群体性事件管理推演与应对措施研究"(项目编号:14CGL050)和中央高校基本科研业务费专项资金"基于复杂网络的微博用户关系发现研究"(项目编号:JBK141106)研究成果之一。

作者简介:王林森(ORCID:0000-0002-1203-9119),讲师,博士,E-mail:wanglinsen1978@126.com;王学义,教授,博士后,博士生导师。

收稿日期:2017-09-07 修回日期:2017-11-15 本文起止页码:71-77 本文责任编辑:王善军

模，为后续算法的展开奠定数据基础。

其次是传导节点与微博系统/外部信息世界的关系模型构建问题。Q. Gao 和 G. Dong 等人根据传导节点的信息主流方向，将其划分为外向型与内向型；并指出传导节点之所以作用巨大，是因为其能够以指向网络的需求作为依据，从内/外网络中主动获取信息[7-8]。G. Maira 等人提出了信息“泵”的概念，指出内向型传导节点成功的要素之一是能够及时而准确地把握所在网络的信息需求[9]。X. Shuai 和 A. J. LAM 等人的研究表明内向型信息传导热点，实质上是能够自动获取内网信息需求与外网高匹配度资源的搜索引擎与广告引擎复合体[10-11]。W. Liao 等人归纳了信息传导热点的运营与盈利模式，并研究了微博类媒体自建信息传导热点的可行性[12]。但上述工作普遍仅发现和解释传导热点现象，缺乏传导热点检测与发现的实用性研究成果。针对上述问题，本研究将以复杂网络节点的信息传导行为建模为基础，研究微博传导热点的行为特征，并规划对应的检测与预测算法。

最后是内向型传导热点检测与预测问题。P. Fornacciari 和 S. Alzahrani 等人研究了微博系统中的内向型传导热点行为特征以及信息扩散途径[13-14]。P. Sibbald 等人基于能量谱理论，研究了微博传导热点的信息扩散特征与活动影响[15]。J. Zhao 等人将微博传导热点视为物理学中的外部热源接触点，研究了传导信息的生灭与覆盖过程[16]。X. Jin 等人将复杂网络理论代入微博系统中加以应用，把复杂网络中的传导节点识别与发现机制，在微博系统中进行了验证[17]。但上述工作中尚缺乏行之有效的传导热点监测与预测算法，而事后热点识别与发现机制对于社交网络管理的辅助决策意义不大。基于上述问题，本研究借助热功率谱算法对上述成果予以修正和补充，从信息热度扩散理论角度出发，研究微博节点传导信息的扩散方式与影响程度，最终实现微博传导热点的预测与发现算法。

2 传导节点与预测流程

2.1 传导节点及相关定义

根据复杂模型传导（关联）节点的原始模型，本研究扩展和定义了微博信息传导热点的基本数据结构，即定义转帖与转载内容超过自身信息量的 75% 的节点为传导节点。在微博信息空间中，设某传导节点 i 的当前位置为 C_i，被其传导的信息特征向量的方向与微博信息空间的基准向量之间存在 θ_i 度的夹角，而当前的 i 节点传导半径 r 是传导节点与传导空间边缘（特征相差最大的另一节点）的最大距离，则传导节点的传导空间可以采用一个五元组来表示，即 $S_i = <\theta_i, x_i, y_i, \alpha, r, >$，在这个五元组中，二元组（$x_i$, y_i）是传导节点 i 在微博信息空间中的位置；而传导节点 i 的传导范围为 Si，此时可以用 $S_i \cup S_j$ 表征传导节点 i 与另一节点 j 的组合传导范围，此时，如设微博空间中存在 N 个传导节点，则该节点群体的组合传导范围为 $\cup_{i=1}^{N} S_i$，此时可以将某个组合传导范围 $\cup_{i=1}^{N} S_i$ 占整个微博传导空间 S_Ω 的比例，当作该节点集合的热度基础指标 f；如果定义其中的信息最大传导半径为常量，此时有：

$$f(\theta_1, \theta_2, \cdots, \theta_i, \cdots, \theta_n) = \frac{\bigcup_{i=1}^{n} S_i}{S_\Omega} \qquad \text{公式(1)}$$

而复杂网络理论表明：子网内的有效信息传导通常不会超过 6 次（小世界模型），因此本算法也将传导半径设置为 6，从而减少动态半径测量带来的计算开销。至此，传导热点的识别问题可以转化为：求节点 i，使其信息特征角度 $\theta(\theta_1, \theta_2, \cdots, \theta_i, \cdots, \theta_n)$ 覆盖范围内，其产生的 f 热度值超过阈值，或在传导节点序列中排名靠前。

2.2 总体流程

传导热点预测的总体流程如图 1 所示：

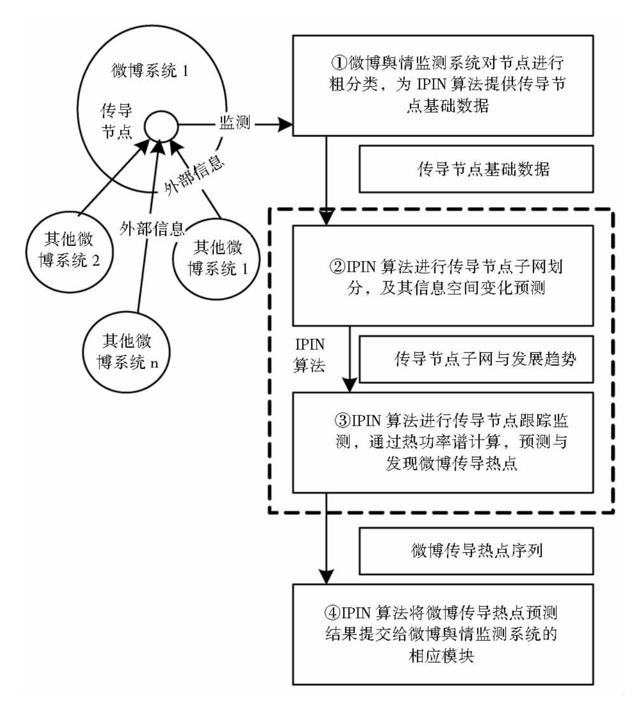


图 1　微博传导热点预测总体流程

步骤 1：在微博节点监测过程中，由监测系统对节点进行总体分类，根据上文定义将传导节点与其他节点进行粗分类，得到传导节点的基础数据。由于该步

王林森，王学义.微博内向型传导热点发现与预测算法研究[J].图书情报工作,2018,62(3):71－77.

骤通常是由微博监测系统自动完成的,因此在下文中不作详述。

步骤2:针对每个监测周期内获取的传导节点信息序列,IPIN算法将通过仿生鸟群(粒子群)算法进行传导节点子网划分,及其信息空间变化预测,具体方法是:基于微博节点的复杂网络属性,发现微博系统中的传导节点的最大容纳数;经过有限次鸟群算法迭代后,将微博信息空间中传导节点子网进行划分,进而获得该子网的发展变化趋势。

步骤3:将经过粗筛和划分的传导节点序列,采用IPIN算法进一步对其进行跟踪预测,具体方法是:将传导节点序列信息片段中的特征能量(传导频次与影响权重)视作热度,通过热功率谱计算,预测其发展变化趋势,进而通过排序和阈值过滤,即可得到微博传导热点子序列。

步骤4:将微博传导热点子序列提交给微博舆情检测系统。

3 算法描述

3.1 变量定义

本研究设定传导的最大距离(跳数)r为6,而新浪与网易微博的数据显示95%以上的传导距离在4跳以内,因此该设定符合国家相关规定中的基本监测指标要求。此外,设定传导节点的信息特征角度数量为α。此时,对整个微博信息空间中的内向型传导节点数量进行评测,基于复杂网络的传导节点理论,结合上文定义,可以得到传导节点的最大容纳数:

$$N \leqslant \frac{\ln(1-p)}{\ln\left(1-\frac{a^2}{2S_\Omega}\times\frac{S_\Omega-\pi^2-r(C_\Omega-2\pi)}{S_\Omega}-\frac{a^2/2}{S_\Omega+rC_\Omega+\pi^2}\times\frac{\pi^2+r(C_\Omega-2\pi)}{S_\Omega}\right.} \qquad 公式(2)$$

在该式中,S_Ω为需要进行传导热点预测与监测的微博信息空间(通常将整个空间且分为若干子空间进行监测),C_Ω是微博信息空间的直径(某空间内关联信息的最大传导距离),而p是当前空间的信息特征密度。

3.2 初步工作

复杂网络中通常用仿生算法进行信息传导节点的变迁分析,最终发现其总体变化趋势,实现热点预测的初期工作,IPIN算法借鉴这一经验,充分利用前期微博舆情监测的数据积累,通过基于仿生鸟群的有限迭代方法来初步发现和预测传导节点子网,并对传导热点进行粗筛。其中的最大迭代次数可以由上文中的传导热点最大容纳数进行控制。初步发现算法的主要思路是:在迭代过程中,将传导节点的信息特征向量的位置信息作为其发展趋势的主要检测对象,通过构建起信息传导活动的主特征方向变化因子来进行其发展方向的初步预测,以判断传导信息的发展方向是否能够引领或符合微博空间的主导信息场。在具体实现步骤中,将传导节点在当前微博信息空间中的信息位置虚拟化,传导节点作为仿生鸟群中的独立飞行的鸟(或粒子),而把整个微博监测空间视为鸟类飞行的虚拟信息场,当鸟(或粒子)在微博信息空间中发生移动或进行信息传导时,其不但受到虚拟信息场中同类信息元素(粒子或鸟)的吸引力,而且受到过分接近或反方向飞行的鸟的排斥力,两种传导力量使得传导节点i的信息位置发生改变,并且最终呈现出一定的总体发展趋势,而这一趋势正是判断传导节点是否能成为传导热点的关键之所在,计算的具体方法为:传导节点(鸟或粒子)的信息位置采用当前信息位置在微博空间内的倾角,传导节点的变化速度可以用其状态位置在微博空间中旋转的角速度标识,而IPIN算法通过鸟(节点)在微博信息空间中的运动轨迹来判断和预测其发展方向的变迁,最终传导节点变化速度以及信息位置可以表述为:

$$v_{in}^{k+1}=w(t)v_{in}^{k}+c_1r_1(p_{bestin}^{k}-x_{in}^{k})+c_2r_2(g_{bestin}^{k}-x_{in}^{k})+c_3r_3\theta_{vf}$$

$$x_{in}^{k+1}=x_{in}^{k}+v_{in}^{k+1} \qquad 公式(3)$$

在公式(3)里,c_3是鸟(传导节点)的加速算子,而r_3为0－1之间的随机变迁系数,θ_{vf}是鸟在微博信息空间中两种合力作用下的变迁角度,其余指标与鸟群算法一致。公式(3)里的θ_{vf}可表述如下:

$$\theta_{vf}=\theta_{max}\times e^{\frac{1}{F_i}} \qquad 公式(4)$$

在公式(4)里,θ_{max}是传导节点单次可能改变的最大角度之值,而F_i是传导节点所在信息位置受到的外力之值。作为公式(3)的补充,鸟(节点)信息状态变化的惯性权重w(t)如下所述:

$$w(t)=w_{max}-(w_{max}-w_{min})\times tan\left(A_w\times\left(1-\frac{t}{t_{max}}\right)^{\delta}\right)+w_{min} \qquad 公式(5)$$

在公式(5)里,t_{max}是传导热点最大容纳数(迭代阈值),而t为是目前所处的迭代次,w_{max}是最大惯性权重,而w_{min}是最小惯性权重,A_w是鸟(传导节点)的加

速算子,δ 是调节算子。由此可得公式(3)中的限制式有:

$$v = \begin{cases} v_{max}, & if\ v \geqslant v_{max} \\ v_{min}, & if\ v \leqslant v_{min} \end{cases}$$

$$x = \begin{cases} x_{max}, & if\ x \geqslant x_{max} \\ x_{min}, & if\ x \leqslant x_{min} \end{cases} \qquad \text{公式(6)}$$

其中,迭代所需的吸引力可由公式(7)进行计算:另一传导节点 t 与节点 i 之间的距离应小于 r,且应属于同一微博信息空间。此时,鸟(节点)i 受到的来自于 t 的引力如下:

$$f_{it} = \begin{cases} \dfrac{k_g}{(x_{ci}-x_t)^2+(v_{ci}-y_t)^2}, & if\ t \in A(\mathrm{i,r})\ and\ t \not\subseteq A_s(\mathrm{i,r}) \\ 0, & otherwise \end{cases}$$

公式(7)

在公式(7)里,k_g 是既定的引力系数,$(\mathrm{x_{ci}}, \mathrm{y_{ci}})$ 是节点 i 当前坐标,(x_t, y_t) 是 t 的当前坐标,$\mathrm{A(i,r)}$ 的含义是以传导 i 为圆心,r 为半径的信息空间,$\mathrm{A_s(i,r)}$ 是传导节点 i 的信息特征子空间。

与之类似的是,迭代所需的排斥力可由公式(8)进行计算:其中另一传导节点 t 与节点 i 的距离应小于 r,且应属于同一微博信息空间。此时,鸟(节点)i 受到的来自于 t 的排斥力如下:

$$F_{ij} = \begin{cases} \dfrac{k_r}{(x_{ci}-x_{cj})^2+(y_{ci}-y_{cj})^2}, & if\ C_j \in A(i,R_2) \\ 0, & otherwise \end{cases}$$

公式(8)

其中的变量解释与公式(7)相同,最终鸟(传导节点)受到的合力可以用公式(9)中的矢量之和进行表达:

$$F_i = \begin{cases} \sum_{k=1}^{m} F_{ik} + \sum_{j=1,i\neq j}^{n} f_{ij}, & m \geqslant 1, n \geqslant 1 \\ 0, & otherwise \end{cases} \qquad \text{公式(9)}$$

至此,通过仿生鸟群算法进行复杂网络的变迁分析,可以通过 N 次迭代之后,根据其变化方向聚类和分离出若干传导节点子网,并能够以这些各传导节点的变迁方向与传导信息特征,为后续的热点预测与发现工作奠定导向基础。

经过上述仿生鸟群算法的处理后,符合微博信息空间总体变化的传导节点将被进一步进行处理,假设微博空间中当前传导的信息内容序列(一个监测周期)为 B,为降低一次性处理开销,将该序列均等分为 n_s 段,并将这些信息片段的特征能量(传导频次与影响权重)实施傅立叶变换,处理式为:

$$S_{CS}(f_k) = \sum_{r=1}^{n_s} X_{CS}^{y}(f_k) X_{CS}^{y*}(f_k)/n_s \qquad \text{公式(10)}$$

在公式(10)里,$X_{CS}^{y}(f_k)$ 与 $X_{CS}^{y*}(f_k)$ 是总体序列 B 的第 r 信息片段的特征傅立叶变换与其共轭值,而 f_k 是该段微博信息中的热点频率,可以表达为:

$$X_{CS}^{y}(f_k) = \sqrt{(S_{x_1 y_{12} x_2}^{r}(f_k) S_{x_2 y_{23} y_3}^{r}(f_k) \cdots S_{x_{B-1} y_{(B-1)} x_B}^{r}(f_k))^{1/(B-1)}}$$

公式(11)

在公式(11)里,$y_{12}, y_{23}, y_{(B-1)B}$ 是相邻两个微博信息片段的相关系数,如果这些信息片段完全独立,则它们之间的相关系数是 1。而 $S_{x_1 y_{12} x_2}^{y}(f_k), S_{x_2 y_{23} x_3}^{y}(f_k), \cdots, S_{x_{B-1} y_{(B-1)B} X_B}^{y}(f_k)$ 是相邻信息片段之间的热传导交互功率谱,可以表述为:

$$S_{x_1 y_{12} x_2}^{y}(f_k) = [X_{y1}(f_k)\gamma_{12} X_2^{y*}(f_k)] \qquad \text{公式(12)}$$

3.3 热点预测与发现

在建立了传导节点的热传导交互功率谱之后,可进行下列微博传导热点预测与发现工作:

步骤 1:通过公式(11)进行处理,对待处理的信息片段 $x(t)$ 实施传导信息热点与热度敏感分量分解,从而取得 n 个热点分量及其对应的变化趋势分量(携带敏感分量)$n_n(t)$;

步骤 2:通过上一步骤中获得敏感分量,以贝叶斯法选取 m 个高频的热度敏感分量($m < < N$),并通过下一步进行复合功率谱计算:

首先,对所选取每个敏感分量所对应的信息序列片段进行傅立叶变换,得到:X_i^j, $i = 1, \cdots, m, j = 1, \cdots, N$;

其次,计算各个热度敏感分量所在片段的相关系数:

$$\gamma_{i(i+1)}^{2}(f_k) = \frac{\left| \sum_{j=1}^{N} S_{x_i x_{i+1}}^{j}(f_k) \right|^2}{\sum_{j=1}^{N} S_{x_i x_i}^{j}(f_k) \sum_{j=1}^{N} S_{x_{i+1} x_i}^{j}(f_k)} \qquad \text{公式(13)}$$

在公式(13)里,$S_{x_i x_{i+1}}^{j}(f_k)$ 可通过公式(14)取得:

$$S_{x_i x_{i+1}}^{j}(f_k) = X_i^j(f_k) X_{i+1}^j(f_k) \qquad \text{公式(14)}$$

在公式(14)式里,$X_i^j(f_k)$ 是当前节点所在的第 i 个传导节点子网总体平均热度的离散余弦变换值。最后,得到热度复合功率谱 $S_{CS}(f_k)$:

$$S_{CS}(f_k) = (\sum_{j=1}^{N} X_1^j(f_k)\gamma_{12}^2 X_2^j(f_k) \cdots X_{m-1}^j(f_k)\gamma_{(m-1)m}^2 X_m^y(f_k))^{1/\mathrm{m}}/\mathrm{N} \qquad \text{公式(15)}$$

步骤 3:基于上述,可得:

$$powr_i = -(\sum_{k=1}^{K} p_i \ln p_i)/\ln K$$

$$p_k = S_{CS}(f_k)/\sum_{k=1}^{K} S_{CS}(f_k) \qquad \text{公式(16)}$$

$$\sum_{k=1}^{K} p_k = 1$$

在公式(16)里,K 是序列片段中的热点数量。至此通过公式(16),可分离出各片段中的传导节点的热度功率变化趋势,将各传导节点的信息热度功率值按全序列进行累加,并进行排序和阈值过滤,即可得到微博传导热点序列。

4 实验结果分析

为了验证 IPIN 算法的实际处理能力,本研究进行了微博数据处理实验。其中,为了证明 IPIN 算法性能的优越性,实验对比算法选定为 SNMA(social networks monitoring algorithm)微博社交热点预测与监控算法,由于该算法并非专门为预测传导热点而设计,因此本实验中对其进行适当改造,使其预测与监控目标专门指向传导节点,而忽略微博内部的原创型热点。

本次实验的硬件环境为 3 台联想 850 服务器,均采用 Intel i5 CPU/16G 内存/8T 硬盘;软件系统的底层采用 Linux 服务器,为了保证全局数据的一致性,采用 NFS(Net File System)实现微博数据的网络无缝共享。本实验采用的基础数据集为新浪微博提供的 2017 年 4 月 1 日至 2017 年 6 月 30 日的 90 天数据集(期间 1 天服务器维护,未提供数据),最终为了突出传导热点的监测与预测效果,通过“新浪大数据在线”选取了信息外部传导最为频繁的政治类与金融类微博数据集作为最终的实验数据集。该数据集中包含数据项超过 9 千万条(含回复),用户数超过 110 万,“新浪大数据在线”最终认定的传导热点为 4 751 个。

为了保证实验的公正与公平,本次实验分别将上述微博数据集严格地按照微博内容生成与提交的时间顺序注入监测系统,而监测系统中一次只运行一种传导热点预测算法,在处理完整个数据集,并保存最终预测结果后,将清空整个系统的缓存,并重新启动,向监测系统中加载另一种传导热点预测算法,重新注入数据集进行处理,最终得到两种算法独立生成的预测结果后,再进行各项指标的对比。本次实验采用了国际公认的三种舆情监测与预测指标对算法的性能进行检验[11]:传导热点预测覆盖率、传导热点预测误报率和预测准确率。传导热点预测覆盖率的定义为被预测算法标识,并且确系传导热点的节点数目在总体被标识的节点中的比例,该定义代表一定监测周期内(本次实验选取的监测周期为 24 小时),预测算法动态的、在该周期内的预测全面程度;传导热点预测误检率的定义可表述为:一定监测周期内,微博传导热点预测过程中,被预测算法标识的、但并未成长为传导热点的数量,在总体被标识的节点中的比例,该定义代表了一定检测周期内预测算法浪费系统资源的程度[7];传导热点预测准确率的定义为被预测算法标识,并且确系传导热点的节点,在实际传导节点中所占的比例,该定义代表一定监测周期内,预测算法动态的、对实际传导热点的预测准确程度。

图 2 显示了两种算法的微博热点预测覆盖率数据,可以看出:在 90 个监测时间窗口里,IPIN 算法的热点预测覆盖率全面超过 SNMA 算法,在中后期的部分监测时间窗口里,该算法的热点预测覆盖率甚至超过 SNMA 算法的覆盖率达到了 25% 以上,其微博热点预测覆盖性能良好;该算法的预测覆盖率波动程度也远小于 SNMA 算法,稳定性较佳。此外,该算法在进行微博热点预测时,覆盖率上升速度较快,从冷启动开始,到加载现场数据进行预测,稳定在较高的热点预测覆盖率的“启动－稳定周期”较短,从而体现出 IPIN 算法良好的数据处理能力。

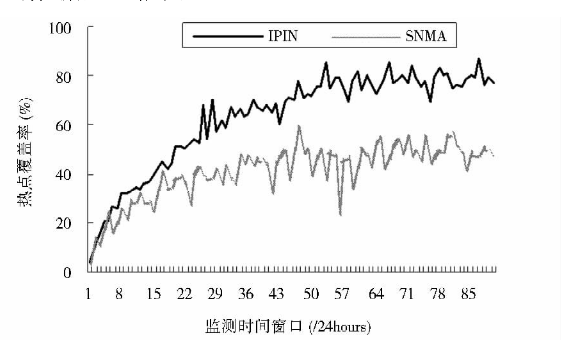


图 2 热点预测覆盖率对比

图 3 显示了两种微博热点预测算法的热点预测误检率数据。可以看出:在 90 个监测时间窗口里,IPIN 算法的误检率总体较低,只有 2 个窗口内的误检率高于 SNMA 算法,而有 25% 以上的窗口内的误检率低于 SNMA 算法的误检率超过 30% 以上,表明了 IPIN 算法具有良好的预测性能,能够为整个监测系统节省大量的计算资源(避免对非传导热点进行持续跟踪)。此外,图 3 中两种算法的误检率数据曲线表明,IPIN 算法的误检率不但回落速度很快,而且进入稳定期后误检率的波动不大,这对于维持整个监测系统的资源开销平稳是非常有利的,体现了较高的性价比。

图 4 显示了两种微博热点预测算法的热点预测准确率数据。可以看出:在 90 个监测时间窗口里,IPIN 算法的准确率总体较高,只有不到 5% 的监测时间窗口里的准确率和 SNMA 算法较为接近,而有 30% 以上的窗口内的准确率高于 SNMA 算法的准确率超过

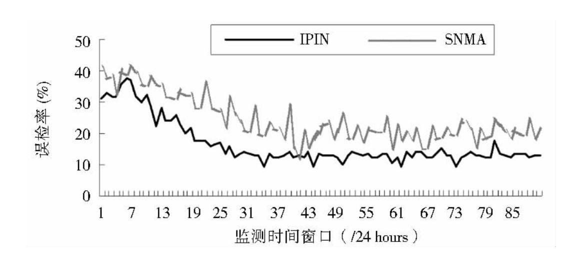


图 3 热点预测误检率对比

30%以上，表明了 IPIN 算法具有良好的预测性能，能够为整个监测工作节省大量甄别工作量。此外，图 4 中两种算法的误检率数据曲线表明，IPIN 算法的准确率不但上升速度很快，进入高位稳定期后准确率的波动不大，体现了较高的算法稳定性。

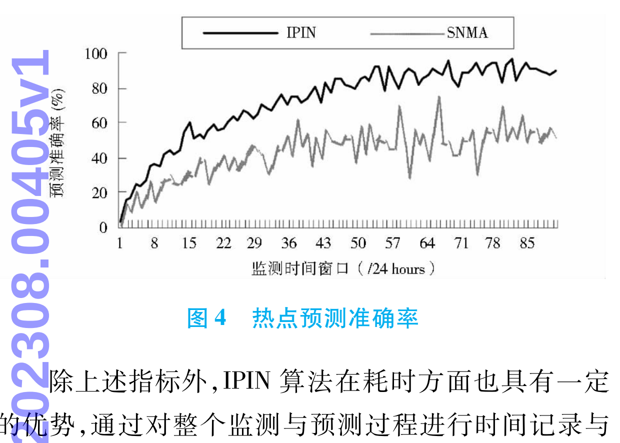


图 4 热点预测准确率

除上述指标外，IPIN 算法在耗时方面也具有一定的优势，通过对整个监测与预测过程进行时间记录与分析发现，该算法的平均热点（单个节点从注册到被确定为热点）发现时耗，较之 SNMA 短 8.57%。其中，有 41%以上的热点发现与预测时耗比 SNMA 算法低 20%以上，这一优势能够为微博监控工作节省宝贵的时间。

5 结论

针对微博传导热点预测与监测中难点与重点问题，笔者提出了一种基于复杂网络分析的预测算法 IPIN。实验证明该算法具有良好的性价比，预测精度和覆盖率较高，且响应速度快，具有一定的实用价值。但限于实验环境与当前的基础理论发展，本研究尚未对外向型信息传导节点进行深入研究，需要在后续工作中深入探讨和分析外向型信息传导节点的作用机制以及监测、预测算法。此外，未来的工作中还将研究传导节点的外部信息关联盲检测、反转载作弊算法等内容。

参考文献：

[1] 吴小兰，章成志. 基于突发事件特征网络的用户社区发现与社区主题演化研究——以新浪微博 H7N9 事件为例[J]. 情报理论与实践，2017，40(5)：94－99.

[2] 刘健，毕强，李瑞. 微博舆情信息传播效果评价指标体系构建研究——基于模糊数据包络分析法[J]. 情报理论与实践，2016，39(12)：31－38.

[3] APARICIO S, VILLAZÓN-TERRAZAS J, ÁLVAREZ G. A model for scale-free networks: application to Twitter[J]. Journal of entropy, 2015, 17(10): 5848－5867.

[4] COLETTO M, LUCCHESE C, ORLANDO S. Twitter for election forecasts: a joint machine learning and complex network approach applied to an Italian case study [J]. Journal of computer networks, 2016, 32(8): 234－238.

[5] MAHARANI W, GOZALI A A. Collaborative social network analysis and content-based approach to improve the marketing strategy of SMEs in Indonesia [C]//International conference on computer science and computational intelligence. New York: Curran Associates Inc, 2015: 373－381.

[6] CHELMIS C. Complex modeling and analysis of workplace collaboration data[J]. Journal of communication, 2017, 14(3): 503－508.

[7] GAO Q, ABEL F, HOUBEN G. A comparative study of users' microblogging behavior on Sina Weibo and Twitter [J]. Journal of computer, 2016, 31(3): 632－644.

[8] DONG G, YANG W. Detecting community pacemakers of burst topic in Twitter [J]. Journal of information systems, 2016, 3(9): 245－255.

[9] MAIRA G, PAULA A A, CLAUDIO P. Large-scale multi-agent-based modeling and simulation of microblogging-based online social network [J]. Journal of communication, 2017, 11(3): 627－638.

[10] SHUAI X, DIN Y, BUSEMEYER J. Modeling indirect influence on Twitter [J]. Journal of social computing, 2016, 8(5): 1－17.

[11] LAM A J. Improving Twitter community detection through contextual sentiment analysis of Tweets [C]//Proceedings of the 54th annual meeting of the Association for Computational Linguistics. Stroudsburg: Valencia Press, 2016: 30－36.

[12] LIAO W. Strategies for spreading information from local to global in social complex networks, cases from a village in China [J]. Journal of computer networks, 2017, 33(2): 12－28.

[13] FORNACCIARI P, MORDONINI M, TOMAUIOLO M. Social network and sentiment analysis on Twitter: towards a combined approach [J]. Journal of multi-media computing, 2016, 3(1): 787－793.

[14] ALZAHRANI S, ALASHRI S, KOPPELA A R. A network-based model for predicting hashtag breakouts in Twitter [J]. Journal of social computing, 2016, 8(11): 423－453.

[15] SIBBALD P. Using social media as a pedagogical tool in graduate public health education and training [J]. Journal of healthcare communications, 2016, 1(2): 4－12.

[16] ZHAO J, WANG L. Research on public opinion propagation in micro-blogging based on epidemic models [J]. Journal of information and management, 2016, 13(2): 235－244.

[17] JIN X, WANG Y. Research on social network structure and public opinions dissemination of micro-blog based on complex network a-

nalysis [J]. Journal of networks, 2015,8(7):1543-1551.

作者贡献说明:

王林森:提出模型结构,采集数据,验证并撰写论文;

王学义:撰写论文,绘制图表。

Research on the Detecting and Prognosis Algorithm of the Micro-blog Hotspot Transmission Inward Node

Wang Linsen[1] Wang Xueyi[2]

[1] Humanlties and Management Sciences School of Southwest Medical University, Luzhou 646000

[2] West China Center for Economic Research Southwestern University of Finance and Economics, Chengdu 611130

Abstract: [**Purpose/significance**] In order to deal with the quick birth-death process and unconspicuous hotspot characteristics of information transmission nodes, a new detecting and prognosis system is proposed based on complex networks. [**Method/process**] In order to deal with those problems, a novel micro-blog information transmission hotspot inward node prognosis algorithm was proposed as the IPIN algorithm based on complex network analysis methods. This paper used this algorithm to build a model with complex network node relations, and found relation sub-networks by related nodes. Then, the thermal power spectrum computing was used to dope out information transmission ranges and prospective effects. [**Result/conclusion**] Data experiment results prove that the IPIN algorithm has higher hot transmission node coverage rate, accuracy rate and better cost-performance than those of the SNMA algorithm.

Keywords: public opinion monitoring micro-blog complex network information transmission node hotspot prognosis

《图书情报工作》2017 年再创佳绩

2017 年,《图书情报工作》在主管主办单位的支持下,在编委会的领导下,在作者、审稿专家、读者和编辑部的共同努力下,期刊在保持良好发展势头基础上,又取得了新的成绩,在相关评价中继续保持不俗的表现:在中国科技信息研究所《中国科技期刊引证报告(2017 年版社会科学卷)》中,《图书情报工作》在情报学学科中综合排名第一,在图书馆学学科中综合排名第二,在"社会科学领域中国科技核心期刊综合评价总分排名"中,位列中国社科 395 种核心期刊第 23 名;在中国知网的"影响力指数"中学科排名第二,连续三年获评"中国最具国际影响力学术期刊";在中国人民大学"复印报刊资料转载指数排名"中,全文转载量继续保持名列本学科第一。据悉,在南京大学 CSSCI 和北京大学《中文核心期刊要目总览》以及中国社会科学院、武汉大学等评价系统中,继续保持良好的地位。

2017 年,《图书情报工作》首次入选《2017 年中国科学院科学出版基金科技期刊排行榜》,并获得中国科学院出版基金资助;首次获得推荐参与申请第三届全国"百强报刊",并最终获得"全国百强科技期刊"称号。

《图书情报工作》旗下的《知识管理论坛》通过国际最重要的开放获取期刊目录 Directory of Open Access Journal (DOAJ)的严格审核,成功入选 DOAJ。由《图书情报工作》发起并牵头的"图情期刊联盟网"沉寂多年,2017 年正式得到中国科学院和文献情报中心的支持,重新启动该项目的研究与试点。

(本刊讯)